DOCUMENT RESUME

ED 185 100 TM 800 105

AUTHOR Haladyna, Tom; Roid, Gale
TITLE The Role of Instructional Sensitivity in the Empirical Review of Criterion-Referenced Test Items.
PUB DATE [Apr 80]
NOTE 29p.; Tables marginally legible. Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Bayesian Statistics; Course Objectives; *Criterion Referenced Tests; Difficulty Level; *Item Analysis; Latent Trait Theory; Models; Pretests Posttests; *Test Construction; *Test Items; Test Theory; *Test Validity
IDENTIFIERS Classical Test Theory; *Discrimination Index; *Instructional Sensitivity; Item Discrimination; Test Analysis

ABSTRACT

An empirical review of test items is described as an essential step in criterion-referenced test development. The concept of test items' instructional sensitivity is introduced, and research is briefly reviewed which describes four theoretical contexts in which instructional sensitivity indexes have been observed: criterion-referenced; classical test theory; Bayesian statistics; and latent trait theory. The study described in this report uses a wide range of item characteristics that transcend the four previously-mentioned approaches; encompasses instructional settings that vary in effectiveness; and has sufficient numbers of items and examinees for each test. The relationship of putative instructional sensitivity indexes to the traditional concepts of item difficulty and item discrimination is explored, using 17 item characteristic indexes. Results strongly suggest that instructional sensitivity is a unique and useful concept in criterion-referenced test development. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185100

THE ROLE OF INSTRUCTIONAL SENSITIVITY IN THE EMPIRICAL REVIEW OF CRITERION-REFERENCED TEST ITEMS

by

Tom Haladyna and Gale Roid

Teaching Research Division

Oregon State System of Higher Education

Monmouth, Oregon

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Tom Haladyna
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running Head: The Role of Instructional Sensitivity

TM800 105

ABSTRACT

The role of empirical item review is described as an essential step in criterion-referenced test development. A test item concept, instructional sensitivity, is introduced, and research reviewed which supports the employment of this concept in item review. Finally, CR achievement test data representing various levels of instructional effectiveness are employed to study the relationship of putative instructional sensitivity indexes to traditional item concepts of difficulty and discrimination. Results strongly suggest that instructional sensitivity is a unique and useful concept in CR test development.

THE ROLE OF INSTRUCTIONAL SENSITIVITY IN THE EMPIRICAL REVIEW OF CRITERION-REFERENCED TEST ITEMS

Many forms of systematic instruction, such as mastery learning (Bloom, 1968), require achievement tests to monitor student progress through a sequence of instructional units. Tests which are expressly designed to measure student progress in systematic instruction have been termed "criterion-referenced (CR)." These CR tests are intended to provide unambiguous information about a student's level of achievement of a single objective, a set of objectives, or a domain of specifiable tasks.

In recent reviews by Millman (1974a), Hambleton, Swaminathan, Algina, and Coulson (1978), and in a paper by Popham (1975), the CR test has been defined as a representative sample of items from a well defined universe of items. The item universe is often referred to as a "domain." Large item domains representing instructional intent are created from which CR tests can be constructed.

It is clear from the domain-based approach to CR testing that the specification of the domain is a necessary and important element. The aim in developing CR tests for systematic instruction is to produce a domain of items that is logically related to the intent of instruction. The

---

Support for this research was obtained through a grant from the National Institute of Education (NIE) Basic Skills Group. The opinions expressed herein do not necessarily reflect the position or policy of NIE, and no endorsement should be inferred.

The authors are indebted to Dr. Craig Scott for his comments on a draft of this manuscript.

essence of the relationship between instruction and testing is captured in a statement by Shoemaker (1975):

> For every instructional program there exists one and only one item universe that is inseparable conceptually from it. (p. 129)

If CR tests are to permit domain-based interpretations of individual performance relevant to a domain, then the way items are created, analyzed and selected for testing has much to do with the reliability and validity of the CR test and its conceptual relationship to instruction.

Therefore, the purpose in this paper is to examine an empirical method by which items may be reviewed. The essence of the empirical method proposed is based on the concept of instructional sensitivity at both the test and item levels. First the background and rationale for this empirical approach to item review is presented. Then a study is presented which provides evidence of the existence of instructional sensitivity as a unique and useful item characteristic to be distinguished from the traditional item characteristics of difficulty and discrimination.

INSTRUCTIONAL SENSITIVITY

Background

The procedures whereby CR tests are constructed to measure individual student progress in systematic instruction are (a) item development, (b) item review, and (c) item selection.

Item development refers to the procedures by which items are generated or written. In most instances, instructional objectives are used to create test items. However, a number a procedures have been devised to create

large and representative groups of items through the implementation of amplified objectives (Popham, 1975), item forms (Hively, Patterson, & Page, 1968), facet theory (Guttman, 1969; Engel & Martuza, 1976; Berk, 1978) or a linguistic-based theory (Bormuth, 1970; Roid & Haladyna, 1978).

Item review may take one of two forms: logical or empirical. Hambleton, et al. (1978) have advocated logical item review which involves content specialists and their observations of item-objective congruence. Such procedures lead to logical evidence of content validity for the CR tests that ensue. Empirical procedures involve collecting item data from field tests and are largely based on the notion that unsuitable items exist and must be detected, revised or removed from the item universe prior to constructing CR tests. There is ample evidence to suggest that regardless of the item development process that is used, flawed items will exist which need attention. Empirical item review is one approach to detecting such flawed items. Both logical and empirical forms of item review should contribute to the establishment of highly internally consistent and content valid CR tests.

Item selection may be based on a logical analysis of which items best reflect instructional intent, or items may be chosen based on their empirical characteristics. Neither approach appears suitable in light of the domain-based interpretation desired from most CR test situations. Since representativeness of the content domain is a primary objective in CR tests, items must be randomly sampled from the item universe. This practice has its roots in classical test theory (Nunnally, 1967; Lord & Novick, 1968) and has been prominent in discussions of domain-based CR testing (Hively,

1974; Millman, 1974a, 1974b; Hambleton, et al., 1978) and has been advocated for generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Rationale

In systematic instruction, test results should reflect what an instructor believes occurs during instruction relevant to the objectives of instruction. Generally, CR test performances for uninstructed students is low prior to instruction and high following instruction. The difference in preinstruction and postinstruction means is an indicator of instructional affectiveness.

Thorndike (in Jackson & Messick, 1967, p. 203) observed: "Each item is in a very real sense a little test all by itself." In other words, what might be true for the test in terms of student responses may also be true for the items that comprise the test. If CR tests of our instructional intent, as operationally defined through the objective or set of objectives upon which the test is based, are sensitive to the effects of instruction, then, analogously, items should also be sensitive to the effects of instruction. Items that are instructionally sensitive may be one of the most appropriate indicators of CR item validity.

Glaser (1963) summarized the intention thus:

> ...achievement tests, used primarily to provide information about differences in treatments, need to be constructed so as to maximize the discriminations made between groups treated differently and to minimize the differences between the individuals in any one group. Such a test will be sensitive to the differences produced by instructional conditions.

> For example, a test designed to demonstrate the effectiveness of instruction would be constructed so that it was generally difficult for those taking it before training and generally easy after training. (p. 376)

In summary, test items that represent our instruction intent should yield test item performances that parallel CR test performance. Items that do not detect differences from pretest to posttest should be reviewed to ascertain if (a) instruction has been faulty or (b) the item is flawed in some way or simply inappropriate.

INSTRUCTIONAL SENSITIVITY: AN EMPIRICAL STUDY

Item Characteristic Indexes

Two traditional item characteristics are difficulty and discrimination. Difficulty is simply the ratio of correct responses to the total responses for any item, while discrimination is the correlation between item and test performance. As described earlier, instructional sensitivity is believed to be a CR test item concept which reflects the role that instruction plays in affecting performances of various groups of examinees at the item levels.

There are four theoretical contexts in which instructional sensitivity indexes have been observed: (a) criterion-referenced, (b) classical, (c) Bayesian, and (d) latent trait. Each of these will be briefly discussed, and relevant research reviewed that integrates the evidence regarding the existence of instructional sensitivity as a unique CR test item characteristic.

Criterion-referenced. The movement away from traditional test practices resulted in the creation of many item characteristic indexes

which were deemed more appropriate for CR test purposes. The first of many indexes to be used to evaluate CR test items was the pre-to-post difference index (PPDI) introduced by Cox and Vargas (1966). Similar indexes were the percent of possible gain (Brennan & Stolurow, 1971) and the Brennan index (Brennan, 1972). Another index introduced in this genre was the phi coefficient (Popham, 1971). All of these items have a common trait, they require uninstructed students' performances. All of these indexes appear to measure instructional sensitivity. Some of these indexes were compared to more traditional discrimination indexes by Hsu (1971), Popham (1971), Helmstadter (1972), Haladyna (1974), Kosecoff and Klein (1974) and Moyer (1977). The results from these studies revealed consistently low relations between these CR instructional sensitivity indexes and the traditional discrimination index computed with a post-instruction sample. In most of these instances there was a high degree of relatedness among these CR indexes. From the studies reviewed here, it is clear that many of these CR indexes come close to representing instructional sensitivity as it was described earlier in this paper, although these relationships appear to vary as a function of the instructional setting and other factors.

Classical. The CR test movement appeared to gain momentum in direct proportion to the expressed dissatisfaction with traditional ways of testing. The classical item discrimination index was rejected, initially due to the suspected lack of variance of CR test scores (Popham & Husek, 1969). More importantly, Millman and Popham (1974) argued that the concepts of item and test variation were unnecessary because the concern in CR testing was for individual responses to a universe of test items and not to relative performances of individuals. The issue of item and test variance has been hotly

debated (e.g., Woodson, 1974a, 1974b; Millman & Popham, 1974; Carver, 1974; Haladyna, 1975). The fact remains that item and test variation does exist with any sensitive measure that is applied to a sample of examinees who differ with respect to their CR achievement. It is how this knowledge is used that is the crux of the issue. Woodson (1974b, p. 139) captured the essence of this position when he stated: "The better the item discriminates among instances of the characteristic with the range of interest, the more information the item gives us."

Despite the fact that traditional item discrimination indexes were not highly related to these CR instructional sensitivity indexes, Helmstadter (1972), Haladyna (1974), and Bernkopf and Bashaw (1976) observed high relationships with these CR indexes when the traditional item discrimination (referred to as "COMPBI") was computed using a sample of instructed and uninstructed students. Further, the relationships between this combined sample discrimination index and the more often used posttest-only sample discrimination index (referred to as "POSTPBI") was not high across varying conditions. Therefore, the heterogeneity of the sample appears to be a key factor in determining the usability of the traditional discrimination index as a measure of the instructional sensitivity of a CR test item.

Bayesian. Three discrimination indexes based on Bayes Theorem were introduced by Helmstadter (1974). All three indexes require collateral information in the form of preinstruction test results, or at least a low-performing group. The indexes are: (a) B1--the probability that a student has knowledge given s(he) gets the item right, (b) B2--the probability the student does not have knowledge given s(he) gets the item wrong, and (c) B3--the probability of making a correct classification, e.g., mastery or non-mastery. In Helmstadter's study, a moderately high correlation was observed

between PPDI and the COMPBI and Bayesian indexes. Haladyna and Roid (1976) detected similar relationships. Clearly, these Bayesian indexes reflect instructional sensitivity, although how these indexes work across a variety of instructional settings is unclear.

Latent trait. Several item characteristics are derivable from the application of the one-parameter latent trait theory (the Rasch model). The item parameter is difficulty, while other item statistics include the mean square fit (MSF), a measure of the item's conformance to a theoretical item characteristic curve, and z-difference (ZDIFF), which is essentially the the normalized difference in item difficulty estimates for instructed and uninstructed samples. ZDIFF appears to function as a type of instructional sensitivity index when applied to instructed and uninstructed samples. Haladyna and Roid (1976) found a high degree of relationship between ZDIFF and instructional sensitivity indexes from other theoretical contexts. There has not been sufficient research on the applicability of latent trait theory on the problem of measuring CR item characteristics, particularly instructional sensitivity.

One study by Crehan (1974) was omitted from this review due to the fact that the methodology differed from other studies. In Crehan's study, items were selected on the basis of their empirical characteristics, and test results were compared in terms of two criteria, CR reliability and validity. The problem with this approach is that this empirical test building procedure destroys any opportunity to achieve domain-based interpretations. Interestingly, the results of the study suggested no significant advantage for any of the empirical item selection procedures employed.

The studies reviewed are limited in a number of ways. First, seldom are more than just a few item characteristics studied. Second, the number

of test items is often too few to provide sufficient information. When correlational procedures are used, the relationships among various indexes is subject to random fluctuation due to the small number of items upon which these correlations are based. Third, samples of examinees are often too small, thus lending to the possibility that estimates of item characteristics are unstable. Finally, instruction may be ineffective, thus nullifying the use of instructional sensitivity indexes. If the CR test employed in such research is not relevant to instruction, then the data obtained and results of the research have very limited usefulness.

Objectives of the Study

The study that follows was designed to avoid some of these shortcomings by (a) examining a wide range of item characteristics that transcend four distinctly different theoretical perspectives in measurement, (b) including instructional settings that vary widely in instructional effectiveness, and (c) insuring sufficient numbers of items and examinees for each CR test being employed.

The specific questions of the study were:

1. Is instructional sensitivity a distinctive item characteristic?
2. If so, which of a large number of these instructional sensitivity indexes spanning four theoretical perspectives is most appropriate?

METHOD

Date Sources

Seven data sets[1] were identified that possessed the following characteristics:

1. Items were administered to instructed and uninstructed students.
2. Each data source (test) had at least 25 items.

3. Items were administered to at least 50 examinees for each of two conditions, instructed and uninstructed.

4. The data sources represented a wide range of possibilities, educational levels and subject matters that exist in systematic instruction.

Descriptive statistics for these data sets are presented in Table 1.

Insert Table 1 about here

The test results shown in Table 1 are arranged in terms of instructional effectiveness, ranging from a high of 55.5% to a low of 9.9%. Sets C and D were from a mastery-based course for college level education students, while the other sets were administered in non-mastery based situations. Data set A was a situation where all possible items in an item universe were administered to the students, each item being generated from one of two item forms.

Analysis of Data

To answer the first of the two research questions posed, item characteristics hypothesized to measure (a) difficulty, (b) discrimination, and (c) instructional sensitivity were computed. The total set of indexes and their abbreviations are given in Table 2.

Insert Table 2 about here

PPDI is the simple difference in item difficulty for instructed and uninstructed students. COMPBI is the point biserial correlation between

---

[1] The data sets in this study were provided by Dr. Kevin Crehan from his 1974 study. His contribution is gratefully acknowledged.

item and test performance for combined samples. PREPBI and POSTPBI are the same index computed for pretest and posttest samples. While PREPBI and POSTPBI are considered traditional discrimination indexes, research by Haladyna and Roid (1976) has indicated that COMPBI is a type of instructional sensitivity index. PREDIFF, POSTDIFF, and COMDIFF are traditional difficulty indexes computed for the pre-instruction, post-instruction, and combined samples.

B1, B2, and B3 are computed in the following manner:

$$B1 = \frac{(POSTDIFF)(COMDIFF)}{(POSTDIFF)(COMDIFF) + (PREDIFF)(1 - COMDIFF)}$$

$$B2 = \frac{(POSTDIFF - PREDIFF)(COMDIFF)}{(POSTDIFF - PREDIFF)(COMDIFF) + (1 - POSTDIFF)(POSTDIFF)}$$

$$B3 = (POSTDIFF - COMDIFF)(COMDIFF)(POSTDIFF) + (POSTDIFF - PREDIFF)$$

Item statistics derived from the Rasch latent trait model were estimated using a program developed originally by Wright (1977)[2]. Z-difference is computed by taking the normalized difference between item difficulty indexes derived from pretest and posttest examinees' responses divided by a standard error of the difference. Rasch item difficulties and mean square fits were estimated from each of these samples and from the combined sample.

While the Rasch model has one item parameter, difficulty, the mean square fit is often used as a discrimination index. High mean square fits usually lead to the rejection of the item from the item domain. As a consequence, the mean square fit was categorized as a discrimination index for the present study.

---

[2] The program was modified and provided by Dr. Fred Forster of the Portland Public Schools. His cooperation and contributions are acknowledged.

RESULTS AND DISCUSSION

For each data set all 17 indexes shown in Table 2 were computed and factor analyzed using a principal components analysis. Factors were labeled using information presented in Table 2 for guidelines. Significance of factor loadings was determined by a procedure described by Gorsuch (1974, p. 185) that employs a variant of Fisher's z-test for the significance of a correlation. Because factor loadings represent the correlation between a variable (in this case an item characteristic index) and a hypothetical factor, the loadings tend to be maximal estimates of the relationships and not subject to the normal tests of significance. Gorsuch suggests that the standard error appropriate for tests of significance of factor loadings should be doubled. Thus, factor loadings were determined using this criterion.

If convergent and discriminant evidence are manifest, then the proportion of significant correlations for instructional sensitivity indexes for the factor of instructional sensitivity should be high, while the proportion for other sets of indexes (i.e., difficulty and discrimination) should be low for that factor. About 71% (30 of 42) of all instructional sensitivity indexes are significantly related to the instructional sensitivity factor, while 31% (13 of 42) of difficulty indexes and only 9% (3 of 35) of discrimination indexes are significantly related to instructional sensitivity.

The same tendencies hold true for difficulty, where 93% (37 of 42) of all difficulty indexes were significantly correlated to a difficulty factor in contrast to 40% (17 of 42) for instructional sensitivity indexes and 9% (3 of 35) for discrimination indexes. For the discrimination factor,

51% (18 of 35) of all discrimination indexes were significantly correlated to the discrimination factor, while only 5% (2 of 42) of the instructional sensitivity indexes and none (0 of 42) of the difficulty indexes were significantly correlated with this discrimination factor.

Thus, the convergent and discriminant evidence in support of instructional sensitivity and difficulty as CR item characteristics is quite sound, while the evidence for discrimination is considerably less supportive. It is clear that there is a certain degree of overlap between instructional sensitivity and difficulty and virtually no overlap between discrimination and the other two item characteristic indexes.

Addressing the second question of this inquiry, the significant loadings for all 17 indexes and seven data sets are presented in Table 3. As shown there, for instructional sensitivity factor loadings for PPDI and ZDIFF were consistently the highest, ranging from .50 to .97. Classical point biserials based on combined samples (COMPBI), as well as all three Bayesian indexes, appeared to function erratically as instructional sensitivity indexes. An analysis of the formulas for the Bayesian indexes reveals that they are greatly affected by the ceiling of item difficulty for the post-instruction sample. In other words, the shift from pre- to post-instruction item difficulties may be small, but the Bayesian indexes will remain high as long as the post-instruction difficulty index is high.

Incidental findings, but ones of interest, are those for difficulty and discrimination. The combined samples difficulty index (COMDIFF) provided the most stable and consistently highest factor loadings for the difficulty factor, ranging in loadings from .75 to .99. In fact, pretest difficulty estimates and posttest estimates of difficulty were fairly accurate in most instances. Rasch difficulty estimates for the three

sample conditions were not as consistent as was expected. Since the Rasch model is said to be independent of the sample with respect to estimating item difficulty, one might expect these factor loadings for the Rasch difficulty estimates to be higher, regardless of the sample used.

Discrimination indexes performed rather poorly in contrast to other item characteristic indexes. No index was particularly effective in measuring discrimination (defined as the relation between item and test performance). Despite the very obvious fact that mean square fits are not, per se, discrimination indexes, there is enough evidence in these data to suggest that the mean square fit is about as good a discrimination index as conventional point biserial indexes for CR tests, although none of these indexes over all conditions performed very well.

CONCLUSIONS

An analysis of the definition of CR testing and its role in systematic instruction suggests that item domains need to be specified representing the instruction intent. The item domains may be created by using any of a wide range of techniques (e.g., linguistic-based theory of item development, item forms, instructional objectives). Following the development of items, a process needs to occur whereby items are judged to be logical members or non-members of the item universe. It seems reasonable and desirable to then conduct field tests to determine if empirical evidence is present to support the keeping of each item in the universe or discarding or revising the item. In some instances the item generating procedure may stand revision. Once items have been logically judged and empirically tested, CR tests are created by sampling items from the item universe using sampling plans, as suggested by Millman (1974b). These procedures should lead to domain interpretations that are commonly desired from CR tests in systematic instruction.

Further, the logical and empirical procedures prescribed insure a high degree of content validity.

With respect to methods for empirically screening items, the concept of instructional sensitivity has been introduced and discussed as it importantly bears on CR testing. The study presented herein revealed (a) that instructional sensitivity is a unique characteristic of CR test items, and (b) that one of the most dependable of these instructional sensitivity indexes is also the simplest in conceptualization and computation, namely, PPDI. Interestingly enough, PPDI was one of the first indexes introduced as a substitute for traditional discrimination. Since its introduction by Cox and Vargas (1966), subsequent research has supported the notion that PPDI best measures instructional sensitivity.

Coupled with logical analysis of items by content specialists, an empirical review of items based on the concept of instructional sensitivity should lead CR test developers to the establishment of item domains which will insure dependable and content valid CR tests.

REFERENCES

BERK, R. A. An application of structural facet theory to objective based achievement testing. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, Virginia, 1978.

BERNKOPF, S., & BASHAW, W. L. An investigation of criterion-referenced tests under different conditions of sample variability and item homogeneity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

BLOOM, B. S. Learning for mastery. Evaluation Comment, 1968, 1, 1-12.

BORMUTH, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

BRENNAN, R. L. A generalized upper-lower item-discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

BRENNAN, R. L., & STOLUROW, L. M. An empirical decision process for formative evaluation. Research Memorandum No. 4. Harvard CAI Laboratory, Cambridge, Mass., 1971.

CARVER, R. P. Two dimensions of tests--Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

COX, R. C., & VARGAS, J. A comparison of item selection techniques for norm-referenced tests. Paper presented at the meeting of the American Educational Research Association, 1966.

CREHAN, K. D. Item analysis for teacher-made mastery test. Journal of Educational Measurement, 1974, 11, 255-262.

CRONBACH, L. J., GLESER, G. C., NANDA, H., & RAJARATNAM, N. The dependability of behavioral measurements. New York: John Wiley, 1972.

ENGEL, J. D., & MARTUZA, V. R. A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the meetings of the American Psychological Association, Washington, D.C., September, 1976.

GLASER, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

GORSUCH, R. L. Factor analysis. Philadelphia: Saunders, 1974.

GUTTMAN, L. Integration of test design and analysis. In Proceedings of the 1969 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1969.

HALADYNA, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

HALADYNA, T. M. On the edumetric-psychometric dimension of tests. American Psychologist, 1975, 30, 603-604.

HALADYNA, T., & ROID, G. The quality of criterion-referenced test items. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

HAMBLETON, R. K., SWAMINATHAN, H., ALGINA, J., & COULSON, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

HELMSTADTER, G. C. Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction. Paper presented at the meeting of the American Psychological Association, Hawaii, 1972.

HELMSTADTER, G. C. A comparison of Bayesian and traditional index of test item effectiveness. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, 1974.

HIVELY, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

HIVELY, W., PATTERSON, H. L., & PAGE, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

HSU, T. Empirical data on criterion-referenced tests. A paper presented at the meeting of the American Educational Research Association, New York, 1971.

KOSECOFF, J. B., & KLEIN, S. P. Instructional sensitivity statistics appropriate for objective-based test items. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, 1974.

LORD, F. M., & NOVICK, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

MILLMAN, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974. (a)

MILLMAN, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14, 17-21. (b)

MILLMAN, J., & POPHAM, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

MOYER, J. E. Alternative reliability indices. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, April, 1977.

NUNNALLY, J. Psychometric theory. New York: McGraw-Hill, 1967.

POPHAM, W. J. (Ed.). Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

POPHAM, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

POPHAM, W. J., & HUSEK, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

ROID, G., & HALADYNA, T. A comparison of several linguistic-based, multiple-choice item writing algorithms. Paper presented at the meetings of the American Educational Research Association, Toronto, March, 1978.

SHOEMAKER, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

THORNDIKE, R. L. The analysis and selection of test items. In D. N. Jackson & S. Messick (Eds.), Problems in human assessment. New York: McGraw-Hill, 1967.

WOODSON, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64. (a)

WOODSON, M. I. C. E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140. (b)

WRIGHT, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

AUTHORS

HALADYNA, TOM. Address: Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361. Title: Associate Research Professor. Degrees: B.A. Illinois State University, M.A. San Jose State University, Ph.D. Arizona State University. Specialization: Educational Measurement, Statistics, and Research Methodology.

ROID, GALE. Address: Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361. Title: Research Professor. Degrees: A.B. Harvard University, M.A., Ph.D. University of Oregon. Specialization: Measurement, Evaluation, Instructional Development.

Table 1

Descriptive Statistics for CR Test Data Used in the Study

Data Set	Test Length	PRETEST N	PRETEST Mean	PRETEST s.d.	POSTTEST N	POSTTEST Mean	POSTTEST s.d.	Difference in Means
A	100	99	18.9%	5.7%	84	74.4%	28.3%	55.5%
B	100	256	32.9	13.7	254	74.8	17.9	41.9
C	45	117	42.0	21.7	110	75.3	5.0	33.3
D	45	65	44.3	8.9	96	69.2	8.6	24.9
E	30	80	43.8	18.4	77	67.3	16.3	23.5
F	26	81	66.2	21.0	81	85.7	11.4	19.5
G	40	81	56.1	17.4	81	66.0	18.6	9.9

Table 2

Item Characteristic Indexes and Their Abbreviations

Theoretical Perspective	Item Characteristic Constructs		
	Instructional Sensitivity	Difficulty	Discrimination
1. Criterion-referenced	PPDI		
2. Classical Theory	COMPBI	COMDIFF PREDIFF POSTDIFF	PREPBI POSTPBI
3. Bayesian	B1 B2 B3		
4. Rasch	ZDIFF	COMRDIFF PRERDIFF POSTRDIFF	COMMSF PREMSF POSTMSF

Table 3

Results[1] of Factor Analysis of 17 Item Characteristics for Each of Seven Data Sets

Data Set:	A	B	C	D	E	F	G	A	B	C	D	E	F	G	A	B	C	D	E	F	G
ISI Indexes	Instructional Sensitivity Factor							Difficulty Factor							Discrimination Factor						
1. PPDI	87	60	88	88	94	97	84		-73												
2. COMPBI	76	45		97	91	75			-86	-70					-51						96
3. BI	98			79	82				-95	-80			98	-99							
4. B2	89	83	92			89					90	91		95							
5. B3	66	91	85			85		73			96	98		96				58			
6. ZDIFF	92	50	92	92	94	90	96		82												
Difficulty Indexes																					
1. COMDIFF	-50	-46						-98	-75	-99	-86	-81	94	96							
2. PREDIFF	72	-84				-60		-85	-43	-86	-99	-98	79	97							
3. POSTDIFF				73	79			-67	-95	-88	-67		96	90							
4. COMRDIFF								96	60	98	86	82	-82	-99							
5. PRERDIFF	51	53				61		82		86	95	98	-77	-96							
6. POSTRDIFF	65			74	-78			64	72	88	65		-96	-98							
Discrimination Indexes																					
1. PREPBI					70	48			59	-71					-54		-63				87
2. POSTBI															-87	-77			-85		84
3. COMMSF									67						57	54	78	69			
4. PREMSF																53	88	83		87	
5. POSTMSF			-60												74	42			71		
Eigenvalue	6.3	3.9	5.2	4.4	3.7	3.8	1.4	5.1	6.3	7.6	7.3	8.6	8.5	3.4	1.8	1.6	1.6	1.2	2.2	1.7	8.8
% of Accounted Variance	37%	23%	30%	26%	21%	22%	9%	30%	37%	45%	43%	51%	50%	20%	10%	9%	10%	7%	13%	10%	51%
Order of Appearance	1	2	2	2	2	2	4	2	1	1	1	1	1	1	3	3	3	4	3	3	2

[1] All entries are factor loadings with decimals omitted, and only significant loadings are reported.